Weighing in on End Weight

Stephanie Shih

Department of Linguistics
Stanford University

Department of Linguistics
University of California, Berkeley

Jason Grafmiller

Department of Linguistics
Stanford University

Linguistic Society of America
85th Annual Meeting
Pittsburgh, Pennsylvania
January 9, 2011

The Principle of End Weight

- "Phrases are presented in order of increasing weight." (Wasow 2002: 3; following Behagel 1909; Quirk et al. 1985)

  (1) peas and carrots > carrots and peas

  (2) the attitude of people who are really into classical music and feel that if it's not seventy-five years old, it hasn't stood the test of time >

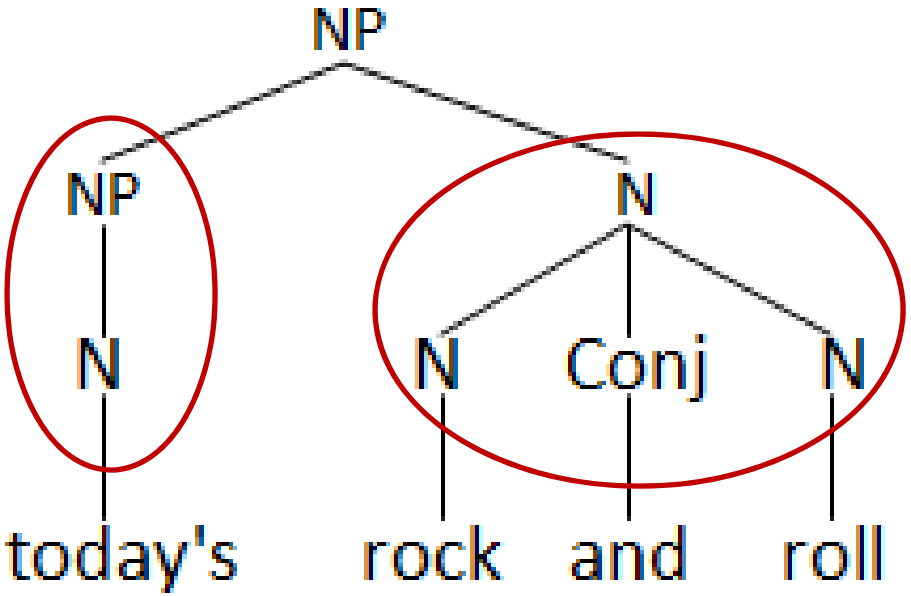
  people who are really into classical music and feel that if it's not seventy-five years old, it hasn't stood the test of time's attitude

- Facilitates planning, production, and parsing
- Cross-linguistic weight at peripheries

What is “weight”?

Syntax

- Syntactic complexity: heavy constituents are structurally more complex.
  - Number of syntactic nodes (e.g., Hawkins 1994)



What is "weight"?
Processing load

- Weight as structural integration cost: heavy constituents require more computational effort
  - Cost of relating an input into a projected structure depends on intervening computations
  - Dependency Locality Theory (Gibson 1998, 2000; Temperley 2007):
  - Each new referent (NP or finite verb) adds to integration cost

What is “weight”?
Phonology

- Phonological complexity: Heavy constituents have complex prosodic properties
  - Number of primary stressed syllables (Anttila et al. 2010; following Selkirk 1984; Zec and Inkelas 1990)

- Phonological weight:
  - Number of syllables (Benor and Levy 2006; McDonald et al. 1993; a.o.)

What is "weight"?

Word Count

- Many studies have used word count as proxy for other weight factors. (e.g., Wasow 2002; Szmrecsányi 2004; Bresnan and Ford 2010; a.o.)

- Correlated with many other measures

Which measure is appropriate?

- Most studies of syntactic alternations focus on syntactic/processing measures of weight
- Influence of phonological weight on syntax less understood
- Multiple weight measures rarely evaluated concurrently on the same data (cf., Szmrecsanyi 2004)

Present Study

The Data

- Two constructions in spoken American English
  (Switchboard Corpus, Godfrey & McDaniels 1992)

(1) Genitive Alternation

- *'s* -genitive ~ *of* genitive
- e.g., the car's wheel ~ the wheel of the car

(2) Dative Alternation

- double object construction ~ prepositional dative (*to*)
- e.g., give the dog the bone ~ give the bone to the dog

Present Study

Weight measures investigated

- Syntactic nodes
- Referents (discourse new)
- Words
- Syllables
- Primary stressed syllables

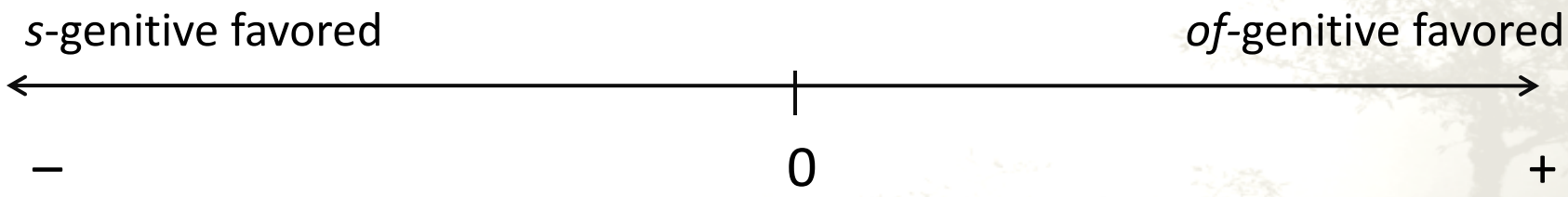
Present Study
Analyses

- Simple and mixed effects regression modeling (Shih et al. 2009; Shih et al. submitted; Hinrichs & Szmrecsányi 2007; Bresnan et al. 2007; Bresnan & Ford 2010; a.o.)
  - 5 individual models using each weight predictor
  - Controlled for other known variables influencing construction choice
  - Model comparison using Akaike Information Criterion (Burnham & Anderson 2004)
- Variable comparison using Random Forests analysis (Strobl et al. 2009b)
  - Single model containing all predictors

Genitives
Fixed Effects Model

- 663 *of*-genitives + 460 *s*-genitives = 1123 total
- Predictors: Possessor animacy, final sibilancy, rhythm (Shih et al. 2009; submitted)
- Comparative weight (Bresnan & Ford 2010)

Comparative weight = log(possessor weight) – log(possessum weight)



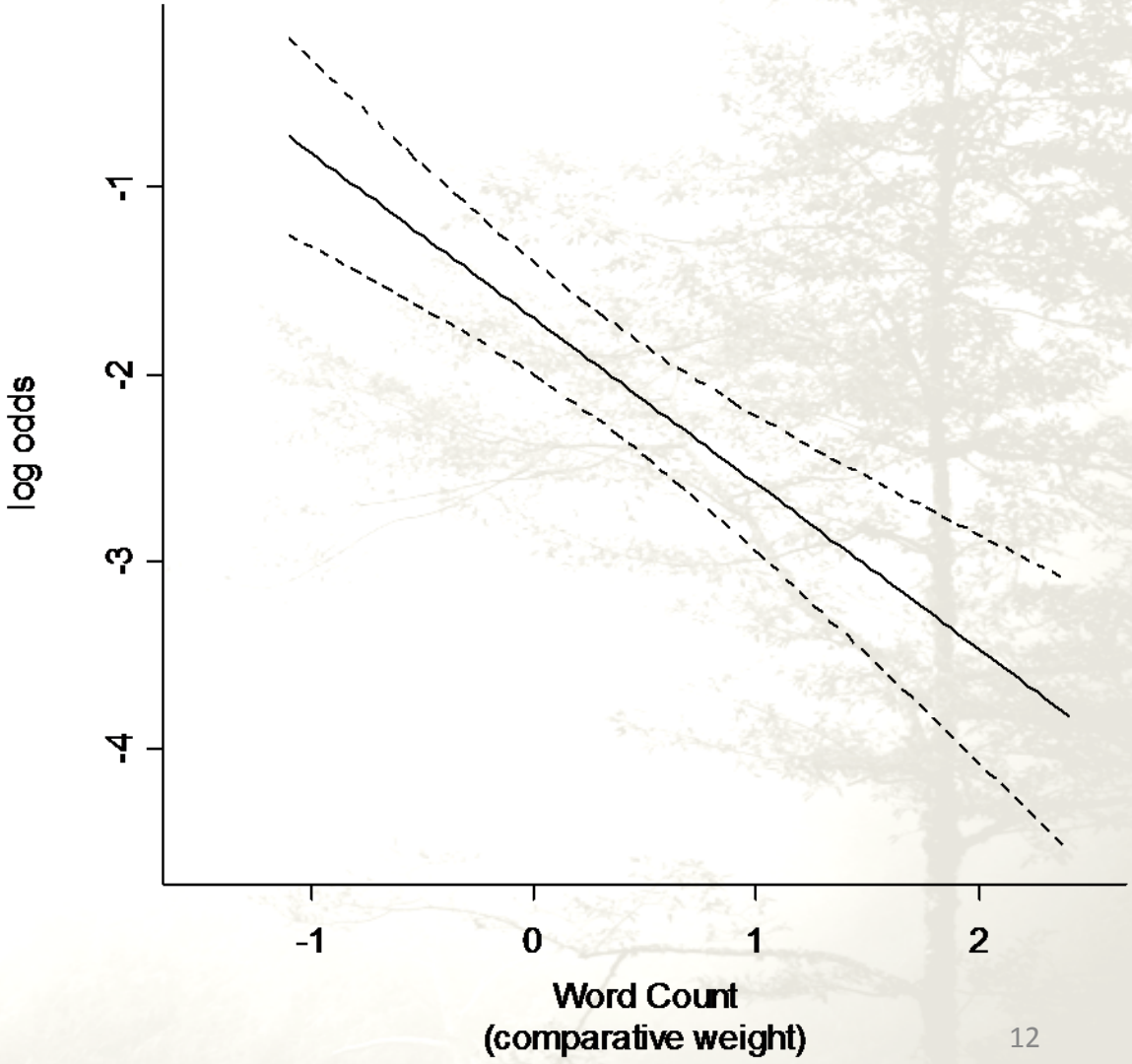
(*Referent counts were not log-transformed.)

Genitives: results

Heavy Possessors favor *of*-gen

- Higher log odds value = higher *s*-genitive likelihood
- Lower log odds value = higher *of*-genitive likelihood

➢ As the number of words in the possessor increases relative to the number of words in the possessum, an *of*-genitive becomes more likely.

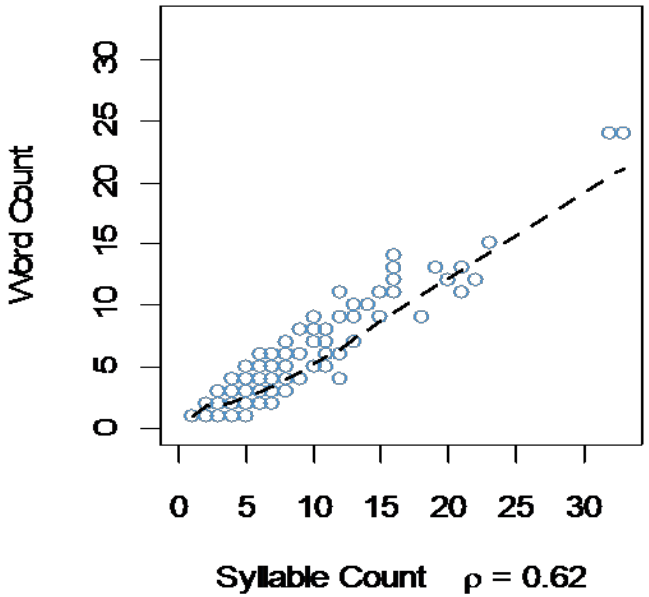


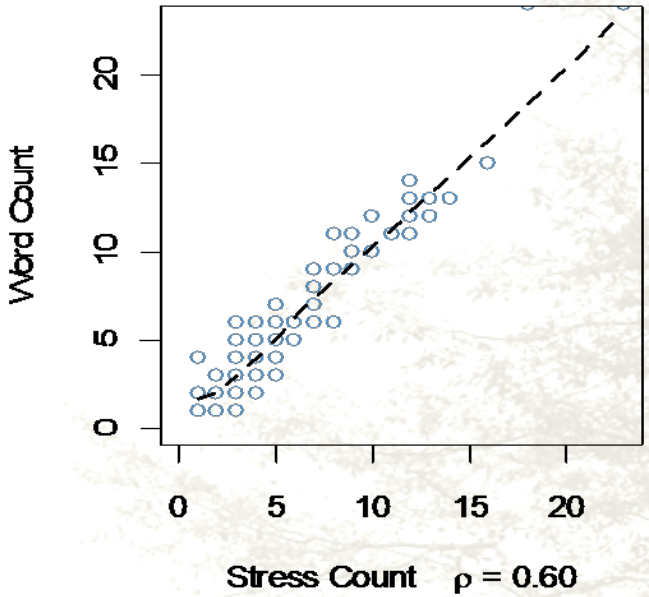
Genitives: results

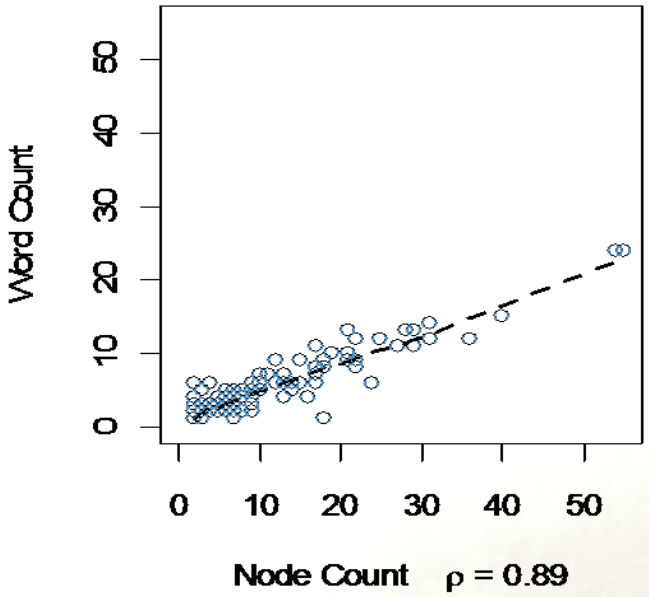
Individual Regression Analysis

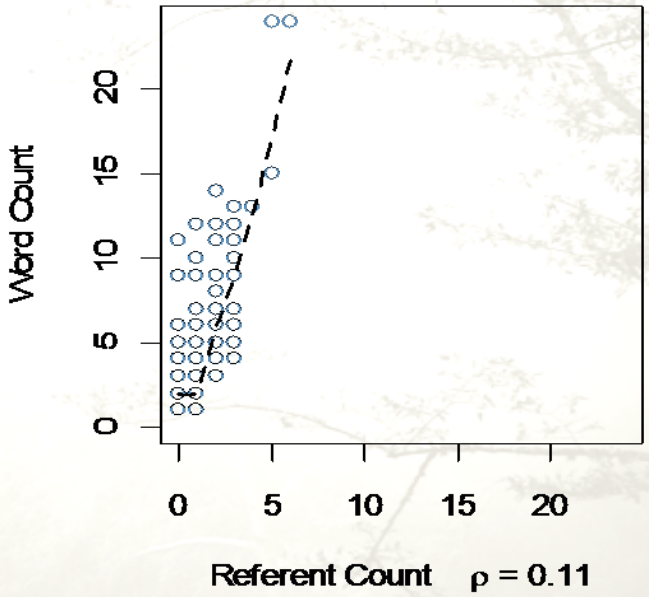
- Nodes
  - β = - 1. 234; z = -6.67; $p < 0.000$ (***)
- Words
  - β = - 0.884; z = -5.50; $p < 0.000$ (***)
- Referents
  - β = - 0.563; z = -3.71; $p < 0.001$ (**)
- Primary Stresses
  - β = - 0.525; z = -3.44; $p < 0. 001$ (**)
- Syllables
  - β = - 0.412; z = -3.42; $p < 0.001$ (**)

Genitives: results
High correlation of factors







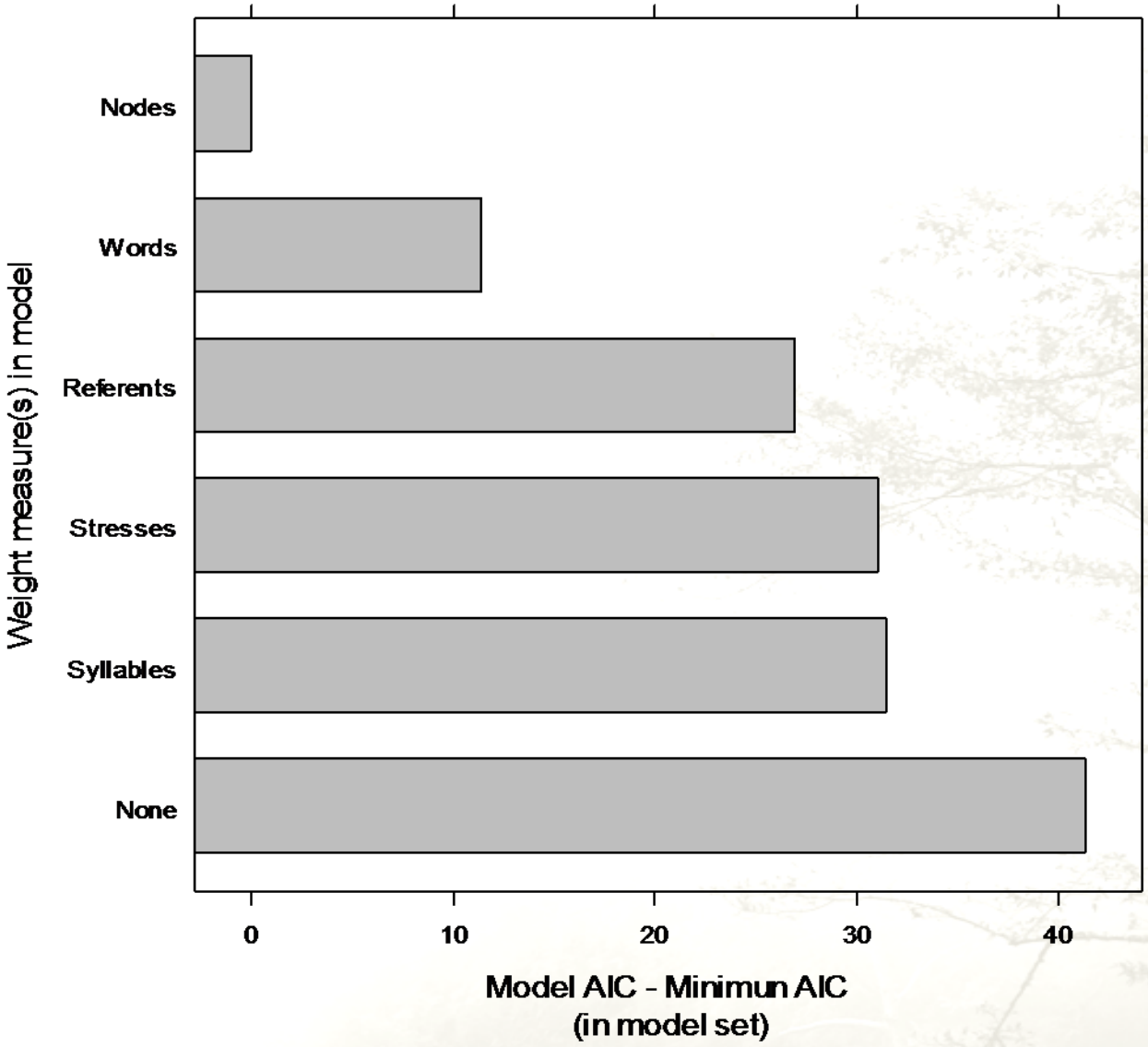


Genitives
Model AICs and factor weights

	Nodes**	Words	Referents	Stresses	Syllables	None
AIC	809.962	821.277	836.889	841.002	841.416	851.218
Δ ($AIC_m - AIC_{min}$)	0.00	11.315	26.927	31.04	31.454	41.256
w_m	0.997	0.003	0.00	0.00	0.00	0.00

- Models with $\Delta < 2$ have substantial support; $\Delta > 10$ have no support
- w_m = the probability that the model is the optimal one in the set (Burnham and Anderson 2006)

Genitives
Comparison of Models

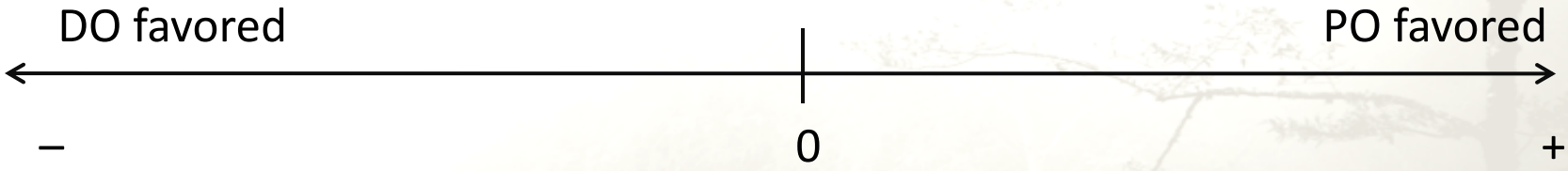


Datives
Mixed Effects Model

- 227 double objects + 183 prepositionals = 410 total

- Mixed effects model (Bresnan et al. 2007; Bresnan and Ford 2010)
  - Fixed effects: animacy of recipient, accessibility of recipient and theme, definiteness of recipient and theme
  - Random effects: Verb

Comparative weight = log(recipient weight) – log(theme weight)

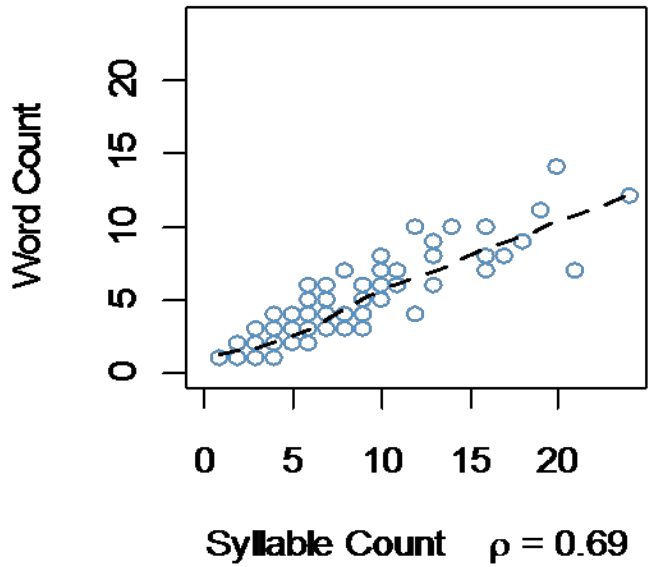


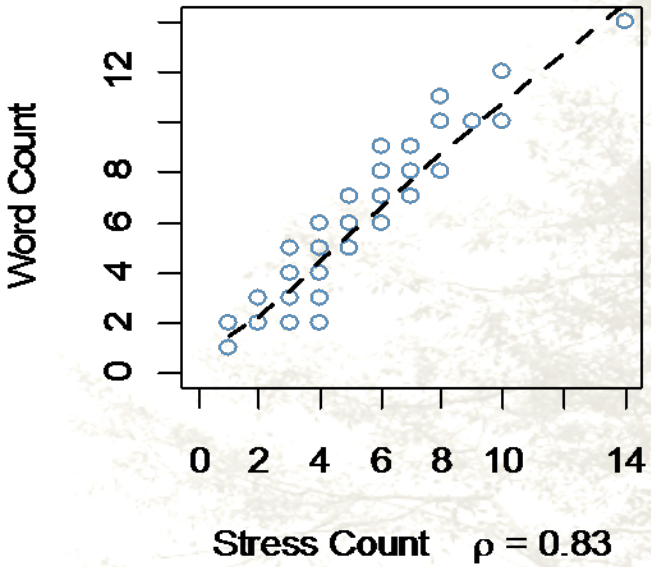
Datives: results

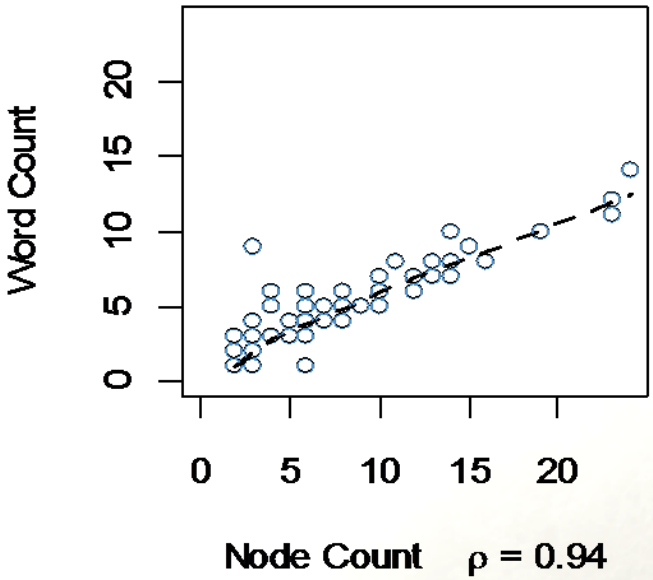
Individual Regression Analysis

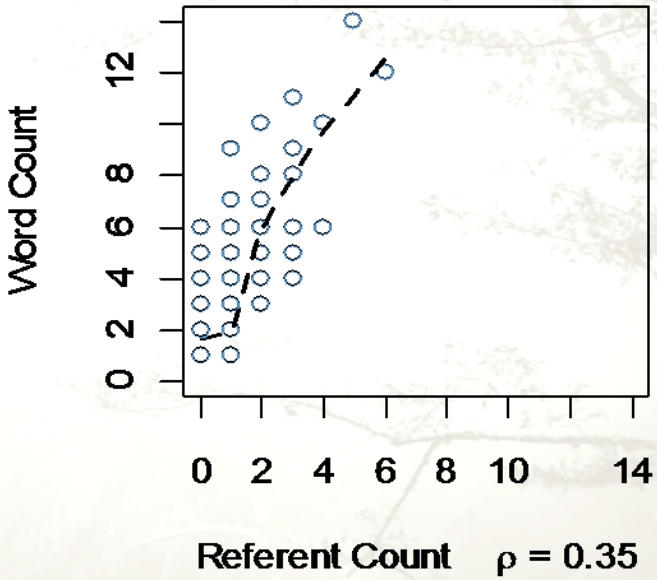
- Nodes
  - β = 1.312; z = 6.685; $p < 0.000$ (***)
- Words
  - β = 1.186; z = 6.877; $p < 0.000$ (***)
- Primary Stresses
  - β = 1.013; z = 6.304; $p < 0.000$ (***)
- Syllables
  - β = 1.040; z = 6.086; $p < 0.000$ (***)
- Referents
  - β = 0.207; z = 1.305; $p = .19$

Datives: results
High correlation of factors







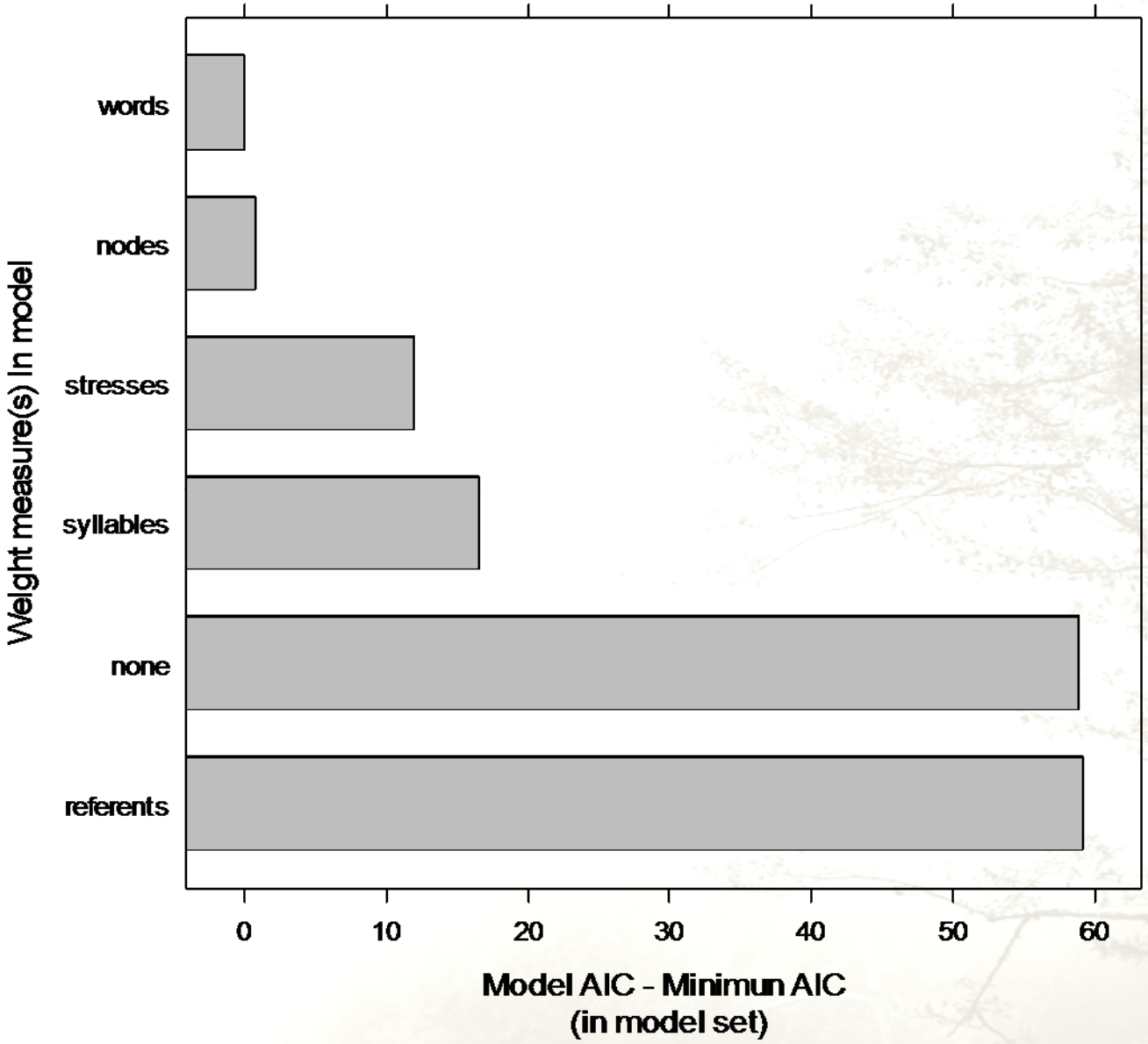


Datives
Model AICs and factor weights

	Words**	Nodes**	Stresses	Syllables	None	Referents
AIC	397.77	398.58	409.81	414.32	456.58	456.90
Δ ($AIC_m - AIC_{min}$)	0.00	0.81	12.04	16.55	58.81	59.13
w_m	0.60	0.40	0.00	0.00	0.00	0.00

- Models with $\Delta < 2$ have substantial support; $\Delta > 10$ have no support
- w_m = the probability that model is the optimal one in the set

Datives
Comparison of Models (AIC)



Random Forests

- Suited to datasets with complex interactions and highly correlated predictor variables (Strobl et al. 2008; 2009a; 2009b; a.o.)
- Recursive partitioning method:
  - Random subsamples of data, each fit with a single classification tree.
  - Randomly restricted set of predictor variables to select from in each split.
- Detects contributions and behavior of predictor variables otherwise masked by competitors.

Random Forests

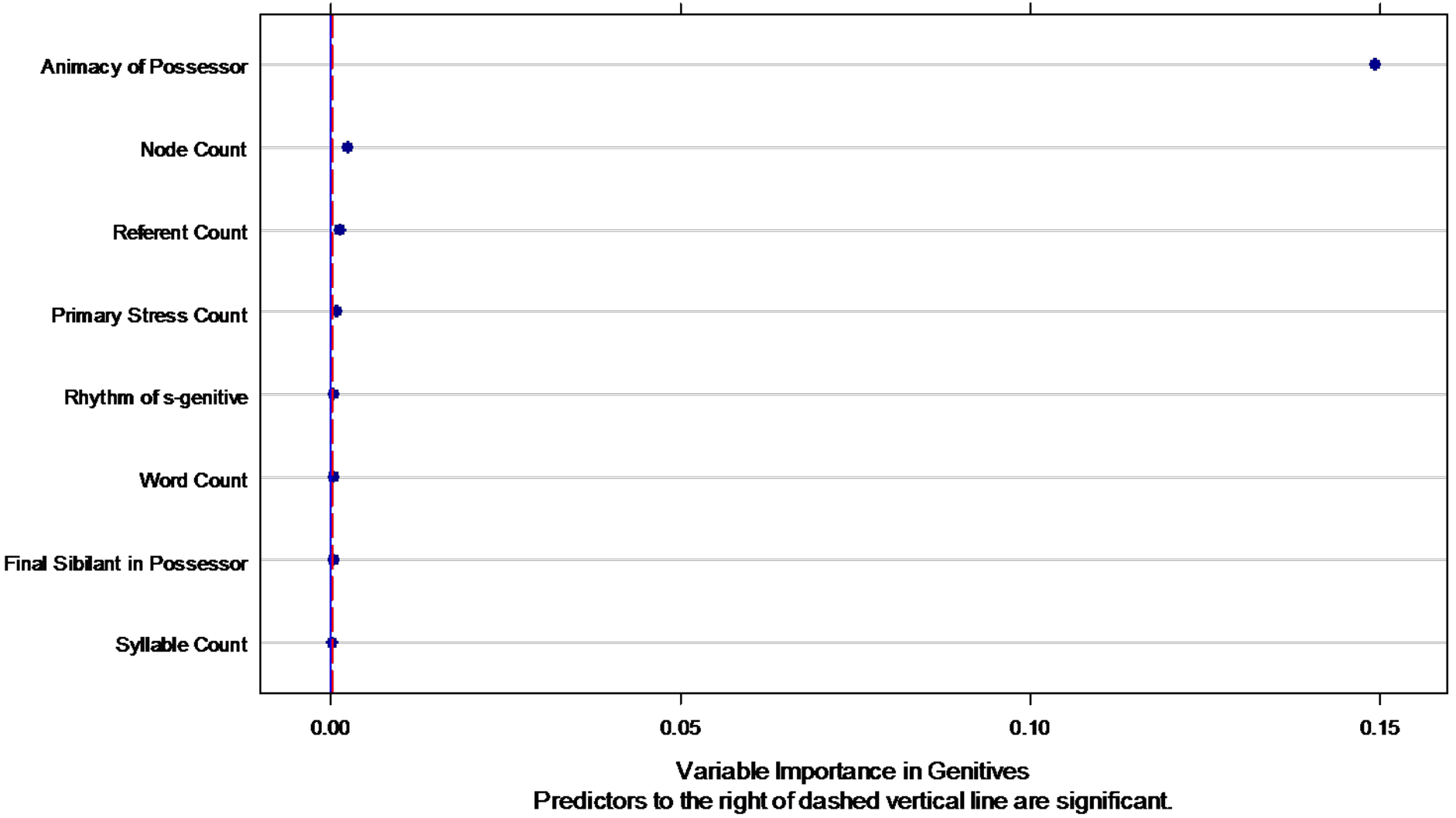
Conditional Variable Importance and Model Parameters

- Conditional Variable Importance
  - Permutation Accuracy: the difference in model accuracy before and after randomly permuting the values of a given independent variable, averaged over all trees in the forest. (Strobl et al. 2009b)
  - Ranks the importance of independent variables.

- Model parameters:
  - Genitives: `ntree` = 2000; `mtry` = 3
  - Datives: `ntree` = 8000; `mtry` = 3
- Model stability verified on two random seeds.

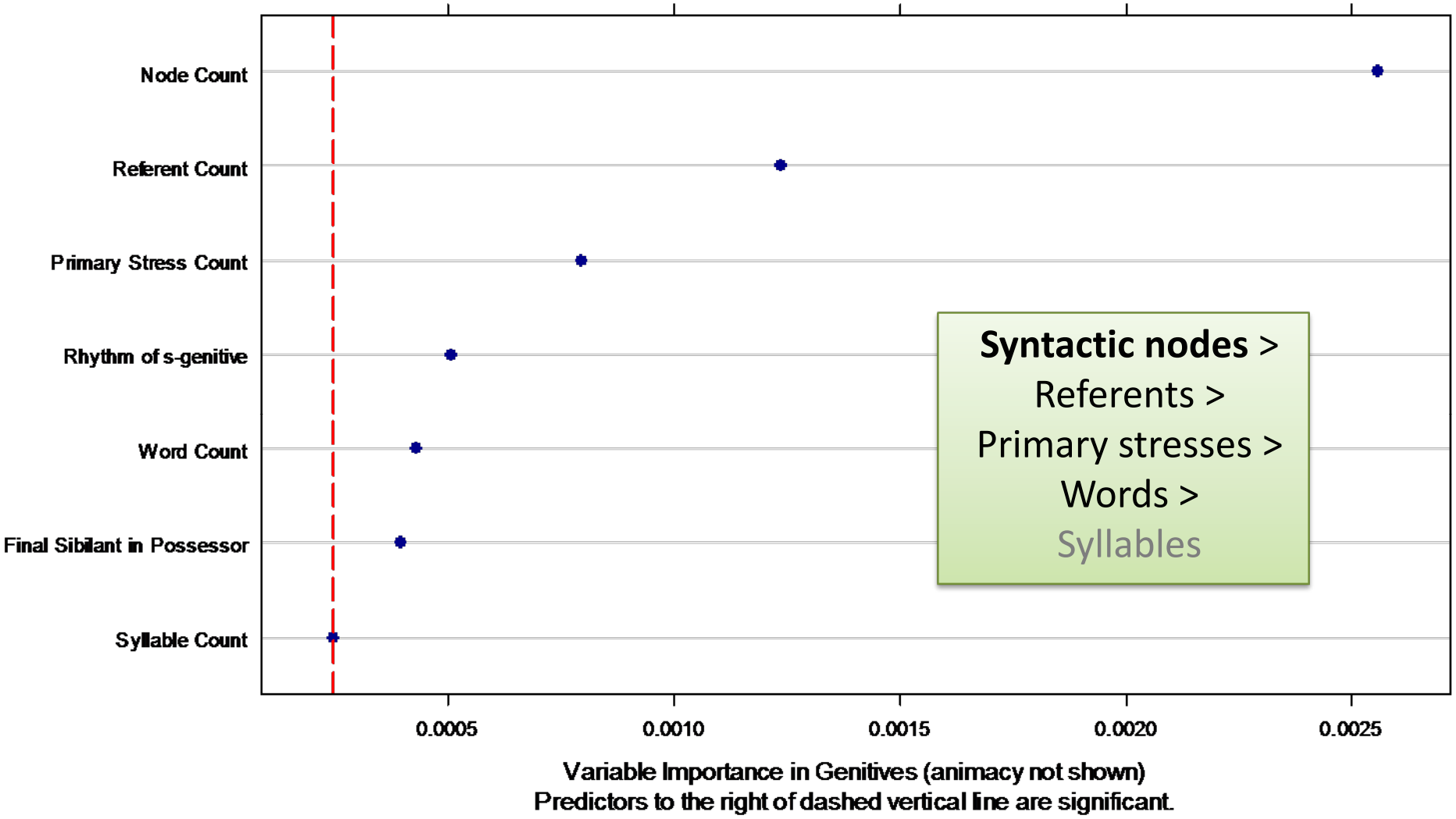
Genitives | Random Forests

Variable Importance



Genitives | Random Forests

Variable Importance

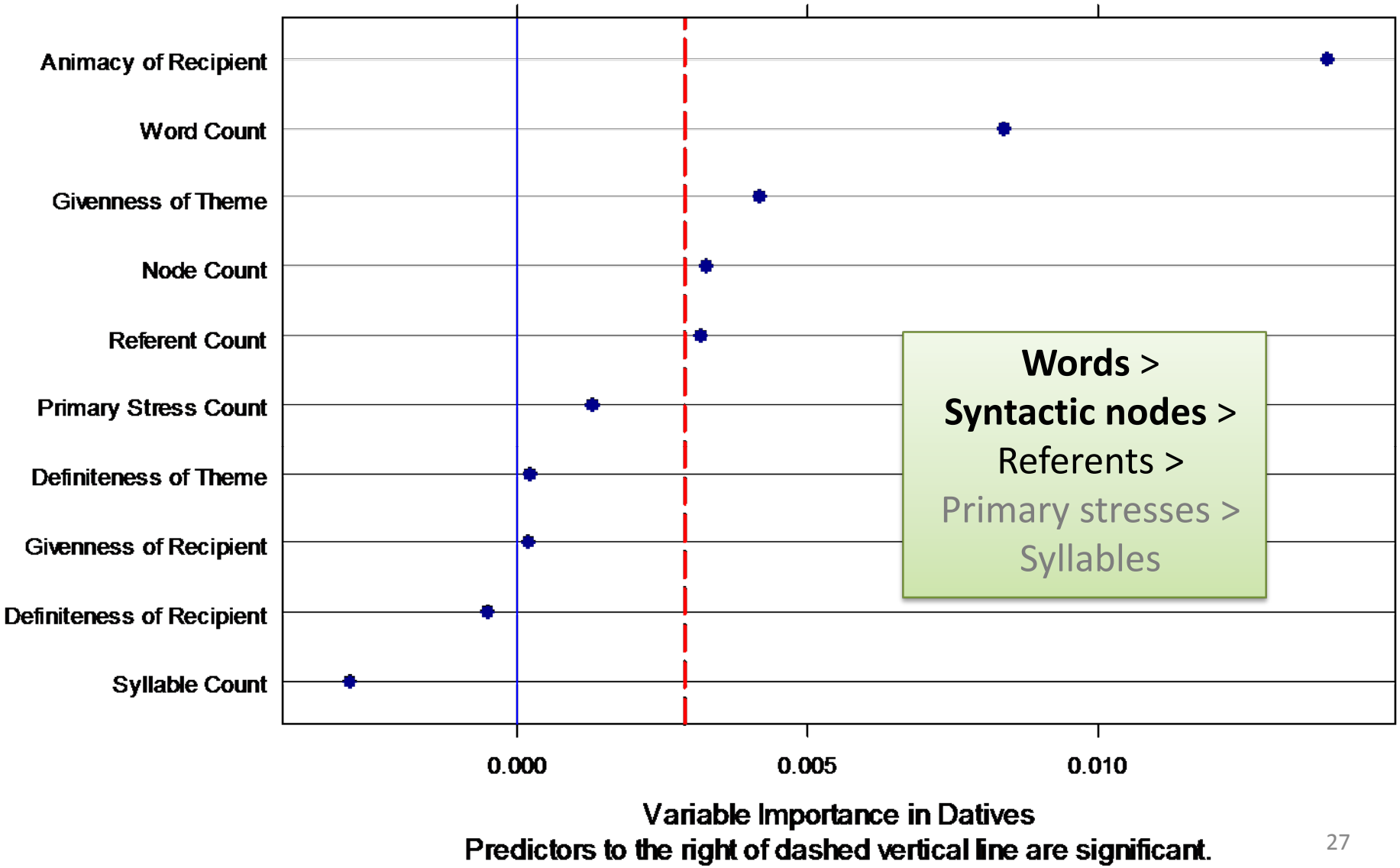


Genitives
AIC vs. Random Forests

	AIC	Random Forest
Genitives	**Syntactic nodes** > Words > Referents > Primary stresses > Syllables	**Syntactic nodes** > Referents > Primary stresses > Words > Syllables

Datives | Random Forests

Variable Importance



Summary
AIC vs. Random Forests

	AIC	Random Forest
Genitives	**Syntactic nodes** > Words > Referents > Primary stresses > Syllables	**Syntactic nodes** > Referents > Primary stresses > Words > Syllables
Datives	**Words** > **Syntactic nodes** > Primary stresses > Syllables> Referents	**Words** > **Syntactic nodes** > Referents > Primary stresses > Syllables

Discussion
Syntactic Complexity

- Number of syntactic nodes = best *individual* predictor of end weight in English genitive and dative construction choice.

- Is "weight" purely syntactic?
  - English binomial ordering studies: number of syllables affect ordering of nouns in binomial pairs. (Wright et al. 2005; cf., McDonald et al. 1993; Benor & Levy 2006)

- At a higher-level domain (i.e., genitives, datives), syntactic complexity is the most salient manifestation of "weight."

Discussion

Word count as a proxy

- Methodologically, the number of words—*though not perfect*—can act as a sufficient proxy for syntactic complexity and 'weight'.

- Dative construction choice:
  - Syntactic nodes and words are the best measures in comparison to the other measures tested.
- Genitive construction choice:
  - AIC: words are second best, though not great.
  - Random forest: not the most important measure

Discussion
Referents and DLT

- In comparison, referents are not the best measures of weight.
  - = Gibson (1998; 2000): Non-given and definite nouns and verbs

- What can contribute to integration costs? (Temperley 2006)

e.g., *the green ball*

Gibson:		x	= 1 new referent
alternatively:	x	x	= 2 new referents

➢ Redefinition of “referents” -> content words?

Discussion
Phonological complexity and weight

- Stresses and syllables rank low as good measures of weight for genitive and dative construction choice.

- Prosodic theory of end weight (=number of primary stresses) is not entirely syntax-independent.
  - phonological words ≈ content words

- Do possible phonetic correlates of weight or complexity play into end weight effects?
  - e.g., duration, complexity of segments, syllable weight or complexity of syllable structure (e.g., Benor & Levy 2006)

Future directions

Weight Beyond English

- How do measures of weight generalize beyond English?

- Is there a better proxy for cross-linguistic syntactic complexity?
  - Morphological complexity and weight?

Conclusion

- Two statistical methods resistant to collinearity:
  - AIC model comparison and selection
  - Random forest conditional variable importance
- Two alternations in spoken American English:
  - Genitives | Datives
- Tested syntactic, processing, and phonological measures of “weight.”
  - Syntactic nodes (syntactic complexity)
  - Referents (Dependency Locality Theory)
  - Words
  - Primary stresses (phonological complexity)
  - Syllables (phonological weight)

Conclusion

- Syntactic-based measures contribute most to weight-driven alternations in higher-level constituent ordering
  - (though perhaps heavily theory dependent)

- Methodologically, the number of words can be an appropriate and sufficient proxy for (syntactic) complexity and weight.

- “Weight” effects cannot be reduced to a single dimension.

Thank you!

Thank you to Arto Anttila for the initial push and subsequent support on this project. Many thanks must go to Joan Bresnan for invaluable advice, discussion, and patience. Acknowledgements also to Matthew Adams, Susanne Gahl, Sharon Inkelas, Victor Kuperman, Beth Levin, Robin Melnick, Daphne Theijssen, Benedikt Szmrecsányi, Tom Wasow, Christoph Wolk, the audience at the *Development of Syntactic Alternations* workshop, and the members of Quorum (UC Berkeley) for further discussion, support, skepticism, questions, and statistical aid.

The authors' names are listed in reverse alphabetical order so as to satisfy the Principle of End Weight.

This material is based in part upon work supported by the National Science Foundation under Grant Number IIS-0624345 to Stanford University for the research project "The Dynamics of Probabilistic Grammar" (PI Joan Bresnan). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Contact:

Stephanie Shih
stephsus@stanford.edu

Jason Grafmiller
jasong1@stanford.edu

Slides available online: http://stanford.edu/~stephsus/ShihGrafmillerLSA2011.pdf

Selected References

Anttila, Arto; Matthew Adams; and Michael Speriosu. 2010. The role of prosody in the English dative alternation. *Language and Cognitive Processes*.

Behagel, O. 1909. Beziehungen zwischen Umfang und Reihenfolge von Satzgliedern. *Indogermanische Forschungen*. 25: 110-142.

Benor, Sarah Bunin and Roger Levy. 2006. The Chicken of the Egg? A Probabilistic Analysis of English Binomials. *Language*. 82(2): 233-278.

Bresnan, Joan; Anna Cueni; Tatiana Nikitina; and R. Harald Baayen. 2007. Predicting the Dative Alternation. in G. Bouma,; I. Kraemer; and J. Zwarts (ed). *Cognitive Foundations of Interpretation*. Royal Netherlands Academy of Science. 69-94.

Bresnan, Joan and Marilyn Ford. 2010. Predicting syntax: Processing dative constructions in American and Australian varieties of English. *Language*. 86(1): 168-213.

Burnham, Kenneth P. and David R. Anderson. 2004. Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociolinguistic Methods Research*. 33: 261-304.

Carnegie Mellon University Pronouncing Dictionary. <http://www.speech.cs.cmu.edu/cgi-bin/cmudict> accessed 2008.

Comrie, Bernard. 2003. On explaining language universals. in M. Tomasello (ed). *The New Pyschology of Language: Cognitive and Functional Approaches to Language Structure*. Mahwah, NJ: Erlbaum. 195-210.

Gibson, Edward. 1998. Linguistic Complexity: locality of syntactic dependencies. *Cognition.* 68: 1-76.

Gibson, Edward. 2000. The dependency locality theory: A distance-based theory of linguistic complexity. in Y. Miyashita; A. Marantz; and W. O'Neil (ed). *Image, Language, Brain*. Cambridge, MA: MIT Press. 95-126.

Godfrey, J. Holliman and J. McDaniel. 1992. SWITCHBOARD: Telephone speech corpus for research and development. *Proceedings of ICASSP-92*. 517-520.

Hawkins, John A. 1994. *A Performance Theory of Order and Constituency*. Cambridge: Cambridge University Press.

Harrell, Frank E, Jr. 2009. Design: Design Package. R package version 2.3-0. http://CRAN.R-project.org/package=Design

Hinrichs, Lars and Benedikt Szmrecsányi. 2007. Recent changes in the function and frequency of standard English genitive constructions: a multivariate analysis of tagged corpora. *English Language and Linguistics*. 11(3): 437-474.

Malkiel, Yakov. 1959. Studies in irreversible conjunctions. *Lingua*. 8: 113-160.

McDonald, Janet L.; Kathryn Bock; and Michael H. Kelly. 1993. Word and World Order: Semantic, Phonological, and Metrical Determinants of Serial Position. *Cognitive Psychology*. 25: 188-230.

Quirk, Randolph; Sidney Greenbaum; Geoffrey Leech; and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. London and New York: Longman.

R Development Core Team. 2010. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <http://www.R-project.org>

Selkirk, Elisabeth O. 1984. *Phonology and Syntax: the Relation between Sound and Structure.* Cambridge, MA: MIT Press.

Selected References (cont.)

Shih, Stephanie. 2010. Random forests for classification trees and categorical dependent variables: an informal quick start R guide. MS. <www.stanford.edu/~stephsus/R-randomforest-guide.pdf>

Shih, Stephanie; Jason Grafmiller; Richard Futrell; and Joan Bresnan. 2009. Rhythm's role in genitive and dative construction choice in spoken English. Paper presented at the 31st annual meeting of the Linguistics Association of Germany (DGfS). University of Osnabrück, Germany. 4 March 2009.

Shih, Stephanie; Jason Grafmiller; Richard Futrell; and Joan Bresnan. submitted. Rhythm's role in predicting genitive construction choice in spoken English.

Strobl, Carolin; Anne-Laure Boulesteix; Thomas Kneib; Thomas Augustin; and Achim Zeileis. 2008. Conditional variable importance for random forests. *BMC Bioinformatics*. 9:307.

Strobl, Carolin; Torsten Hothorn; and Achim Zeileis. 2009. Party on! A new, conditional variable-importance measure for random forests available in party package. *The R Journal*. 1/2: 14-17.

Strobl, Carolin; James Malley; and Gerhard Tutz. 2009. An Introduction to Recursive Partitioning: Rationale, Application, and Characteristics of Classification and Regression Trees, Bagging, and Random Forests. *Psychological Methods*. 14(4): 323-348.

Szmrecsányi, Benedikt. 2004. On operationalizing syntactic complexity. *Journées internationals d'Analyse statistique des Données Textualles*. 7: 1031-1038.

Szmrecsányi, Benedikt. 2008. Probabilistic determinants of genitive variation in spoken and written English: a multivariate comparison across time, space, and genres. in T. Nevalamen; I. Taavitsamen; P. Pahta; and M. Korhonen (ed). *The Dynamics of Linguistic Variation*: *Corpus Evidence on English Past and Present*. Amsterdam: Benjamins.

Temperley, David. 2006. Minimization of dependency length in written English. *Cognition*. 105: 300-333.

Wasow, Tom. 2002. *Postverbal Behavior*. Stanford, CA: CSLI Publications.

Wright, Saundra K.; Jennifer Hay; and Tessa Bent. 2005. Ladies first? Phonology, frequency, and the naming conspiracy. *Linguistics* 43(3): 531-561.

Zec, Draga and Sharon Inkelas. 1990. Prosodically Constrained Syntax. in S. Inkelas and D. Zec (ed). *The Phonology-Syntax Connection*. Stanford, CA: Center for the Study of Language and Information.